Standards for Machine-Readable Course Descriptions

Yannik Langer[1], Carsten Lanquillon[2], Seline Löwe[3], Tom Ochs[4], Elias Ottenbacher[5] and Joshua Zürn[6]

[1]Hochschule Heilbronn, Heilbronn, Germany, yannik.langer@hs-heilbronn.de
[2]Hochschule Heilbronn, Heilbronn, Germany, carsten.lanquillon@hs-heilbronn.de
[3]Hochschule Heilbronn, Heilbronn, Germany, seline.loewe@hs-heilbronn.de
[4]Hochschule Heilbronn, Heilbronn, Germany, toochs@stud.hs-heilbronn.de
[5]Hochschule Heilbronn, Heilbronn, Germany, eottenbach@stud.hs-heilbronn.de
[6]Hochschule Heilbronn, Heilbronn, Germany, jozuern@stud.hs-heilbronn.de

Abstract

Module handbooks at German higher education institutions are predominantly maintained as unstructured PDF or Word documents. The absence of a common, machine-readable format blocks reliable processing by AI-based services and turns cross-institutional course comparison – central to exchange programmes such as EUonAIR – into a manual and error-prone task. A systematic literature review shows that no established standard (IEEE LOM, ECTS, ELMO, schema.org/Course) combines pedagogical depth with machine-readability: cognitive-level coding via Bloom's Taxonomy, Constructive Alignment and NLP processability remain unaddressed. Following the Design Science Research methodology, this study develops a best-practice standard for course descriptions that is both human- and machine-readable (PDF and JSON-LD) and encodes learning outcomes by Bloom level. The artifact is evaluated through expert interviews (Mayring qualitative content analysis, criteria K1–K8) and an exploratory controlled experiment in which two large language models process the same module in free-text and in structured form. Structuring raises extraction F1 from 74 % to 94 %, Bloom-classification accuracy from 64 % to 86 % and lowers the hallucination rate from 27 % to 7 %, at the cost of roughly double the input tokens. Because the experiment covers a single module with one run per model–format–task combination, and because the Bloom level is stated explicitly in the structured input but only implicitly in the free text, these figures should be read as the direction and consistency of an effect rather than as precise performance estimates. The results confirm the value of structured, pedagogically rich course descriptions while identifying authoring effort as the main adoption barrier.

Keywords: *Machine-readable course descriptions; Design Science Research; Bloom's Taxonomy; Constructive Alignment; JSON-LD;*

1. Introduction

This section motivates the project. It describes how course descriptions are currently handled at higher education institutions, derives the research problem and questions, and outlines the structure of the paper.

1.1 Initial Situation

Module handbooks are the central document through which higher education institutions (HEIs) communicate the content, learning outcomes, workload and assessment of their courses. At German HEIs these handbooks are almost exclusively produced and stored as unstructured PDF or Word documents, and every faculty applies its own layout and terminology. No common, machine-readable format exists across institutions [1], [2].

This situation has become a bottleneck for two developments that increasingly rely on structured course data. First, AI-based student services – advising chatbots, recommender systems and automatic question generation – must extract structured information from these documents; when applied to free text, large language models (LLMs) show low extraction quality and a high tendency to hallucinate [2], [3]. Second, international exchange programmes such as EUonAIR depend on comparing courses across partner institutions. Today this comparison and the resulting credit recognition are performed manually, because the underlying descriptions are neither structurally consistent nor interoperable [4], [5].

1.2 Problem Statement and Motivation

Several metadata standards for educational resources already exist (IEEE LOM, ECTS Users' Guide, ELMO, schema.org/Course). However, a systematic review of the literature reveals a structural gap: none of these standards combines pedagogical depth with machine-readability. In particular, the cognitive level of learning outcomes (Bloom's Taxonomy), the coupling of outcomes, teaching methods and assessment (Constructive Alignment, Biggs 1996 [8]), and the suitability of descriptions for natural-language processing are not jointly operationalised in any single standard [1], [6], [7]. As a consequence, learning outcomes remain semantically "empty" for machines, and the comparability promised by frameworks such as ECTS is not realised in practice.

This project addresses the gap by designing a best-practice standard for course descriptions that is simultaneously human-readable and machine-readable and that explicitly encodes pedagogical information. The work is guided by the following research questions:

- RQ1 – Which mandatory fields must a course description contain in order to be both pedagogically meaningful and reliably machine-processable?

- RQ2 – How can pedagogical concepts – Bloom's Taxonomy (C1) and Constructive Alignment (C2) – be operationalised within a machine-readable standard?
- RQ3 – Does structuring a course description according to the proposed standard measurably improve LLM-based extraction and Bloom classification and reduce hallucination compared with free text (B3)?
- RQ4 – How do practitioners assess the practicality and adoptability of the standard relative to existing standards?

The overarching goal is therefore a validated, dual-readable standard that closes the three identified research gaps C1, C2 and B3 and offers HEIs and exchange networks a concrete orientation for producing comparable, AI-ready course descriptions.

1.3 Structure

Section 2 establishes the conceptual foundations: the role of module handbooks, the relevant metadata standards, and the pedagogical concepts of Bloom's Taxonomy and Constructive Alignment, concluding with the related-work gap. Section 3 describes the methodology, which follows the Design Science Research (DSR) process and combines a systematic literature review, expert interviews analysed via qualitative content analysis, and an artifact evaluation. Section 4 presents the results: the designed artifact, the expert evaluation along criteria K1–K8, and the controlled AI experiment. Section 5 synthesises the findings, positions them against the literature, derives recommendations for action and discusses limitations. Section 6 concludes and outlines future work.

2. Fundamentals

This section establishes the conceptual basis of the work – the role of module handbooks and existing metadata standards, the pedagogical concepts of Bloom's Taxonomy and Constructive Alignment, and the combined research gap that motivates the artifact.

2.1 Module Handbooks and Existing Metadata Standards

A module handbook documents, per module, the administrative metadata (name, identifier, credits, workload, responsible person, language), the intended learning outcomes, the content, the teaching and learning methods, and the assessment format. Several standards address parts of this information. IEEE LOM 1484.12.1 offers a mature metadata schema with around 80 fields and broad international adoption, but it relies on verbose XML, carries no pedagogical layer and does not encode the cognitive level of outcomes [6]. The ECTS

Users' Guide anchors workload and credits within the Bologna framework, yet it is purely administrative and does not prescribe a machine-readable format [5]. ELMO v1.0 supports the exchange of transcripts within the EMREX network but is result-focused rather than description-focused. schema.org/Course is web-native and JSON-LD-ready with broad reach, but its vocabulary is shallow and lacks a structured representation of learning outcomes.

Standard	Strengths	What is missing for course comparison
IEEE LOM 1484	~80 metadata fields; technically mature; internationally recognised	No Bloom (C1); no Constructive Alignment (C2); verbose XML, NLP-heavy
ECTS Users' Guide	Bologna context; workload concept; established Europe-wide	Purely administrative; no Bloom; no alignment; no machine-readable format
ELMO v1.0	Transcript exchange; EMREX network; machine-readable	Result-focused, not description-focused; no Bloom; no alignment
schema.org/ Course	Web standard; JSON-LD; actively maintained	High-level vocabulary; no pedagogical depth; no HE-specific fields
Böhm (2021) / SSD	JSON-LD + dual readability; HE context; closest precursor	C1 left open; no formal ECTS/LOM mapping; no empirical evaluation (B3)

Table 1. Existing standards and the gaps relevant to this project (based on 19, 23, 15, 3].

2.2 Bloom´s Taxonomy and Cognitive-Level Coding (C1)

Bloom's revised Taxonomy classifies learning outcomes by cognitive demand into six ascending levels: remember, understand, apply, analyse, evaluate and create. The level is typically expressed through the action verb of an outcome (e.g. "name" → remember, "model" → create). Recent work shows that this classification can be automated with high accuracy – Li et al. (2022) report an F1 of up to 0.95 using BERT-based classifiers, and further studies confirm the feasibility of mapping outcomes to Bloom levels [9], [10]. Despite this technical maturity, a survey of educational ontologies found that none of ten analysed schemas encodes the cognitive level [1]. Cognitive-level coding therefore constitutes the first research gap (C1).

2.3 Constructive Alignment (C2)

Constructive Alignment [8] is the pedagogical principle that intended learning outcomes, teaching and learning activities, and assessment must be coherently coupled. Ramesh et al. (2016) [11] demonstrate that linking outcomes to assessment in a domain ontology is achievable with high accuracy, and Liu et al. (2024) [12] show empirically that generative-AI applications surface the alignment gap. Nevertheless, no established standard formally operationalises the coupling of outcome, method and assessment, which forms the second research gap (C2).

2.4 Machine-Readability and NLP Processability (B3)

Machine-readability requires that a description be parseable into discrete, typed fields rather than interpreted from prose. Text-mining studies on syllabi report solid but limited accuracy when working on unstructured input – for example 83 % with LSA-based methods [4] and improvements with cosine-similarity and Word Mover's Distance approaches [13], [14]. For LLMs specifically, free text yields low extraction quality and elevated hallucination [2], [3]. The absence of any standard that is explicitly designed for NLP processability is the third research gap (B3).

2.5 Related Work and Combined Gap

Concept-matrix analysis (Webster & Watson 2002 [15]) across the reviewed literature shows that the three gaps are addressed in isolation but never jointly. The closest precursor, Böhm (2021), proposes semantically enriched curricula in JSON-LD with dual readability and is the only source close to the project's scope; however, it leaves Bloom coding (C1) and Constructive Alignment (C2) open and provides no empirical evaluation (B3). The central contribution of this work is therefore to combine C1, C2 and B3 in a single, evaluated standard with a formal mapping to existing schemas – a combination that the concept matrix confirms is absent from prior work.

3. Methodology

The project follows the Design Science Research (DSR) methodology of Peffers et al. (2007) [16], which structures the development and evaluation of an artifact into six activities: problem identification, definition of objectives, design and development, demonstration, evaluation and communication. DSR is appropriate here because the central output is a designed artifact – a course-description standard – whose utility must be demonstrated and evaluated rather than merely described. Three complementary methods feed into this process and provide the evidence base for the artifact.

3.1 Systematic Literature Review

To ground the problem definition and derive design requirements for the artifact, a structured, concept-centric literature review was conducted following Webster and Watson (2002) [15].

The search was conducted in three scholarly databases: IEEE Xplore, ACM Digital Library and Google Scholar. Four Boolean search strings were used, covering four thematic areas: machine-readable course descriptions, NLP-based extraction of learning outcomes, educational metadata standards, and course comparability in the Bologna/ECTS context.

Across the twelve database-search-string combinations, the search returned 40,726 records in total. After applying the year filter, 27,192 records remained. Because Google Scholar and some ACM searches returned very large relevance-ranked result sets, the manually screened pool was restricted to the top-ranked records per search string, as documented in the screening log. In total, approximately 295 records were manually reviewed at title level. Title screening reduced the pool to 170 records, abstract screening reduced it to 73 records, and 33 records were retained at the search-run level.

The retained records were exported to Zotero where possible and consolidated manually. Duplicate records and overlaps between search strings and databases were removed before the final eligibility assessment. Based on this consolidation, 20 unique full-text records were assessed for eligibility, of which 17 studies were included in the final concept matrix. The included studies were coded along the three gap dimensions derived from the research questions: cognitive-level coding via Bloom's Taxonomy (C1), Constructive Alignment between outcomes, teaching methods and assessment (C2), and machine-readability/NLP processability (B3).

Inclusion criteria were: (1) the publication addresses course descriptions, syllabi, module handbooks, learning outcomes, educational metadata, curriculum comparability or credit recognition; (2) the publication is relevant to higher education or transferable to higher education contexts; (3) the publication contributes to at least one of the three gap dimensions C1, C2 or B3; and (4) the publication is available in English or German. Exclusion criteria were: (1) publications outside the educational metadata, curriculum or learning-outcome context; (2) publications without relevance to course-level descriptions or comparability; (3) duplicates; and (4) records for which no sufficient abstract or full text was available for eligibility assessment.

The review results were organised in a concept matrix. Each included study was assessed according to whether it addressed Bloom-level coding, Constructive Alignment, machine-readable representation, NLP processability, interoperability with existing standards, and empirical evaluation. This concept-centric synthesis provided the evidence base for the research gap and informed the mandatory-field catalogue of the designed artifact.

3.2 Expert Interviews and Qualitative Content Analysis

To complement the literature review with a practitioner perspective, three semi-structured expert interviews were conducted with stakeholders from Heilbronn University. The interviewees were selected through purposive sampling because their roles directly relate to module handbook creation, study-programme management or credit recognition.

Each interview lasted approximately 30 minutes and covered the current handling of module handbooks, practical problems in course documentation, assessment of alternative standards and perceived implementation barriers. The interviews were analysed using qualitative content analysis based on Mayring (2015). The coding scheme was structured around the research questions and the three gap dimensions C1, C2 and B3.

In addition, each expert assessed four objects – the HHN status quo, the proposed best-practice standard, schema.org/Course and IEEE LOM using eight criteria: completeness (K1), clarity for human readers (K2), pedagogical depth (K3), AI/NLP processability (K4), cross-institutional comparability (K5), effort for authors (K6), alignment with existing standards (K7), and realistic implementability at HHN (K8). Each criterion was rated from 1 = very weak to 5 = very strong. The scores reported in the results section represent the arithmetic mean of the three expert ratings.

3.3 Artifact Evaluation

The artifact was evaluated through an analytical and an experimental strand following Hevner et al. (2004) [18]. The analytical evaluation assessed whether the proposed standard fulfilled the requirements derived from the literature review and expert interviews, especially completeness, clarity, pedagogical depth, machine-readability, interoperability and implementability.

1. The experimental evaluation tested whether structuring a course description according to the proposed standard improves machine processing compared with less structured formats. For this purpose, the same real HHN module, "Einführung in die Wirtschaftsinformatik" (module no. A2.01.01 / 282113; 7.5 ECTS), was prepared in four input formats: Free text – the original unstructured module description as it appears in the HHN module handbook.

2. schema.org/Course – the same content expressed as a JSON-LD record using the schema.org/Course vocabulary.

3. IEEE LOM – the same content expressed in the IEEE LOM 1484.12.1 XML format.

4. Best-practice JSON-LD – the same content encoded according to the proposed three-layer standard with Bloom levels and Constructive Alignment links.

The content basis was kept constant across all formats so that differences in model performance could be attributed primarily to the structure and semantic richness of the input format. Two state-of-the-art large language models were used: Anthropic Claude Sonnet 4.6 and OpenAI GPT-5.5. Each model was presented with identical prompts across three experimental tasks:

1. Information extraction – extracting administrative and pedagogical course information (ECTS, learning outcomes, exam format, prerequisites, teaching methods).

2. Bloom classification – classifying each learning outcome according to Bloom's Taxonomy levels (1–6).

3. EUonAIR advising – answering course-comparison and suitability questions relevant to EUonAIR-style credit recognition.

The outputs were evaluated against a manually defined ground truth. The ground truth was derived from the original HHN module handbook: the administrative and content fields were taken directly from the handbook, and the Bloom level of each learning outcome was coded manually by the authors from the action verb of the original outcome statement, following the classification scheme introduced in Section 2.2. The coding was fixed before the model outputs were scored, and the same answer key was applied unchanged to all four input formats. It must nevertheless be noted that this coding also informed the bloomLevel fields of the best-practice format, so that the target level is stated explicitly in the structured input while it has to be inferred from the free-text and IEEE LOM inputs. The Bloom metric therefore compares a retrieval task against an inference task; the resulting validity threat is discussed in Section 3.4. Three metrics were used:

1. Extraction F1: the harmonic mean of precision and recall for discrete information fields correctly identified in the model output.

2. Bloom-classification accuracy: the proportion of learning outcomes assigned the correct Bloom level (1–6) compared to the expert-coded ground truth.

3. Hallucination rate: the share of factual assertions in the model output that could not be traced to any field in the input document or the ground truth. A response segment was coded as hallucinated if it contained a factual claim about the module (e.g. a credit value, an assessment form, a learning outcome) that was neither present in the input nor derivable from it.

Input-token consumption was recorded as a cost indicator to quantify the trade-off between structural richness and computational expense.

3.4 Methodological Limitations

The methodology has several limitations. First, the literature review was systematic and documented, but the very large result sets in Google Scholar required a restriction to top-ranked records. Second, the expert evaluation is based on a small purposive sample from one institution and is therefore indicative rather than representative. Third, the AI-based evaluation uses one module and two selected LLMs, meaning that absolute performance values may vary with other modules, prompts or model versions. The results should therefore be interpreted as evidence of practical utility within the evaluated setting, not as generalisable performance estimates.

Fourth, and specific to the Bloom metric, the ground truth and the bloomLevel fields of the best-practice format rest on the same manual action-verb coding. Part of the observed gain in Bloom-classification accuracy may therefore reflect the models reading a level that is already present in the input rather than deriving it independently. A leakage-controlled re-analysis – restricting the metric to those outcomes whose cognitive level is not stated explicitly in any input format – was not possible within the scope of this study, because the experiment was not designed to separate the two conditions; it is reported as a priority for future work in Section 6. The extraction and hallucination metrics are not affected by this threat, since their answer key is taken from the original handbook only.

Fifth, each model–format–task combination was executed exactly once. No repeated runs were performed, so no variance, confidence interval or significance test can be reported, and the effect of sampling non-determinism on the individual percentages is unknown. Together with the single-module design, this means that the reported values should be read as an indication of the direction and consistency of the effect across formats, not as measurements of the magnitude of that effect.

4. Results

This section reports the three outcomes of the design and evaluation process: the designed best-practice artifact, its assessment by experts along criteria K1–K8, and the controlled AI experiment comparing free-text with structured input.

4.1 The Designed Artifact

The artifact is a best-practice standard for course descriptions organised in three layers. Layer 1 is a mandatory-field catalogue (module name and identifier, ECTS credits and workload, learning outcomes, EQF/DQR level, prerequisites, language, teaching and exam forms). Layer 2 adds the pedagogical encoding: every learning outcome carries a Bloom level derived from its action verb and a competency category, which also makes the outcome–method–assessment coupling explicit (C1 and C2). Layer 3 ensures interoperability by expressing the description in JSON-LD on the basis of schema.org/Course, with mappings towards IEEE LOM and ECTS. The standard is dual-readable: the same content is rendered as a one-page human-readable PDF and as a machine-readable JSON-LD record.

The standard was instantiated on the compulsory Heilbronn University module “Einführung in die Wirtschaftsinformatik” (module no. A2.01.01 / 282113; 7.5 ECTS; 225 h workload; 6.0 SWS; EQF/DQR level 6; module coordinator Prof. Dr. Philipp Küller). For this module, all six learning outcomes were rewritten as Bloom-anchored statements with an attached competency category, demonstrating that an existing module handbook can be transformed into the proposed format without loss of information.

4.2 Expert Evaluation (K1-K8)

Across the interviews, three cross-cutting findings emerged. Module handbooks are maintained by module coordinators and consolidated by study-programme management, with no formal quality assurance beyond completeness checks; system-related issues create substantial maintenance effort. Compact, one-page formats and ease of maintenance were consistently regarded as critical for adoption, while students rarely consult handbooks because of their complexity. From the internationalisation perspective, recognition relies primarily on external handbooks and is

hindered by inconsistent documentation and ECTS practice; centralised ownership was seen as a way to improve consistency. The comparative scores along K1–K8 are summarised in Table 2.

Criterion (1–5)	HHN status quo	Best practice	schema.org/ Course	IEEE LOM
K1 Completeness	3	4	2	3
K2 Clarity for humans	3	5	3	2
K3 Pedagogical depth	2	4	2	2
K4 AI / NLP processability	2	4	4	4
K5 Cross-institutional comparability	2	3	3	3
K6 Effort for authors	3	4	3	2
K7 Alignment with existing standards	3	3	4	4
K8 Realistic implementability at HHN	4	3	2	2

Table 2. Expert evaluation of four objects across criteria K1–K8 (mean of three interviews, scale 1–5).

The best practice leads on the human and pedagogical criteria – completeness (K1 = 4), clarity (K2 = 5) and pedagogical depth (K3 = 4) – and at least matches schema.org/Course and IEEE LOM on machine-readability (K4 = 4) and comparability (K5 = 3). It trails only on realistic short-term implementability at HHN (K8 = 3), where the established status quo still leads, reflecting the additional authoring effort of a richer format.

4.3 AI-Based Evaluation

In the controlled experiment, structuring the same module markedly improved all three target metrics. The experiment was conducted with Anthropic Claude Sonnet 4.6 and OpenAI GPT-5.5 across the three tasks described in Section 3.3. Each model–format–task combination was run once with identical prompts, and results were averaged across the two models.

Averaged over both models, extraction F1 rose from 74 % (free text) to 94 % (best practice), Bloom-classification accuracy from 64 % to 86 %, and the hallucination rate fell from 27 % to 7 %. The overall index is computed as the arithmetic mean of extraction F1, Bloom accuracy and faithfulness (defined as 1 − hallucination rate), scaled to a 0–100 range. Per-format extraction F1 (×100) was 76/72 for free text, 89/87 for schema.org, 54/50 for IEEE LOM and 95/93 for the best practice (Claude / GPT respectively). IEEE LOM performed worst because it offers no native learning-outcome field and its verbose XML

confused the models. Aggregated into the overall index, the formats ranked: IEEE LOM 58, free text 70, schema.org 85 and best practice 91.

Format	Extraction F1	Bloom accuracy	Hallucination rate	Overall index	Input tokens / call
Free text (control)	74 %	64 %	27 %	70	1 649
IEEE LOM	52 %	56 %	34 %	58	1 561
schema.org/Course	88 %	79 %	12 %	85	2 244
Best practice (exp.)	94 %	86 %	7 %	91	3 415

Table 3. AI-based evaluation across formats (extraction F1, Bloom accuracy, hallucination rate and overall index as means over Claude Sonnet 4.6 and GPT-5.5; input tokens per call as a cost indicator).

The improvement carries a cost: the best-practice JSON-LD consumed roughly twice the input tokens of the free-text version (3 415 vs. 1 649 per call). Structured, Bloom-coded JSON-LD is thus the most reliably usable format for AI processing, but it is also the most token-intensive.

The three metrics should not be read as equally strong evidence. Extraction F1 and the hallucination rate are scored against an answer key taken from the original module handbook and are therefore independent of the format under test. The Bloom metric is not: in the best-practice format the cognitive level is an explicit field, whereas in free text and IEEE LOM it must be inferred from the action verb. The gain of +22 pp thus quantifies the benefit of making the level machine-readable, but it does not isolate how much of that gain reflects improved classification as opposed to successful retrieval of a pre-coded value. Since each combination was run once on a single module, the individual percentages carry no error estimate; what the experiment supports is the consistent ordering of the four formats across both models and all three metrics, not the exact size of the differences.

IEEE LOM's poor performance deserves particular attention. Despite being the most established metadata standard in the educational technology community, its verbose XML structure and the absence of a dedicated learning-outcome field meant that both LLMs struggled to locate and extract the relevant information, frequently misinterpreting XML tags as content. This produced both the lowest extraction F1 (52 %) and the highest hallucination rate (34 %) of any format tested.

5. Findings and Recommendations for Action

The two evaluation strands converge. The AI experiment confirms the value of structure on the machine side: the best practice wins on every metric (extraction F1 +20 pp, Bloom +22 pp, hallucination −20 pp), directly addressing all three gaps – C1 through machine-readable Bloom levels that raise classification quality, C2 through the now-extractable outcome–method–assessment linkage, and B3 through fewer invented details in advising. The interviews confirm the value on the human side, where the best practice scores highest on clarity (K2 = 5) and pedagogical depth (K3 = 4). The findings also position the artifact precisely against the literature: no prior source jointly addresses C1 + C2 + B3 with a formal standard mapping, and the closest precursor [7] leaves exactly the elements this artifact contributes.

There is, however, one consistent trade-off. Practitioners weight authoring effort and realistic short-term implementability (K8 = 3, below the status quo at K8 = 4), and the structured format costs about twice the input tokens. The machine-side gains therefore carry an adoption cost that must be managed rather than ignored.

From these findings the following recommendations for action are derived for HEIs and exchange networks:

- Adopt the mandatory-field catalogue as a minimum schema and treat optional fields as a clearly separated extension layer, so that authoring stays manageable.
- Maintain dual readability: generate the human-readable one-page PDF and the machine-readable JSON-LD from a single source to avoid duplicate maintenance.
- Encode learning outcomes by Bloom level anchored to the action verb of the original outcome, which keeps the coding transparent and auto-classifiable.
- Centralise ownership of module handbooks and add a lightweight quality-assurance step beyond completeness checks, as the interviews identified maintenance and consistency as the key practical risks.
- Pilot the standard with two to three EUonAIR partner institutions to validate cross-institutional comparability and recognition workflows in practice.

Limitations. The evidence base is intentionally focused and should be read with several constraints in mind. The interview sample is small (n = 3) and limited to a single institution, so the practitioner perspective is indicative rather than representative. The AI experiment rests on a single module handbook with one run per model–format–task combination, which yields point values without any estimate of run-to-run variance and makes the absolute percentages dependent on the chosen module, models and prompts. In addition, the Bloom metric is favoured by the best-practice format by construction, because that format states the cognitive level explicitly while the free-text and IEEE LOM conditions require it to be inferred (see Section 3.4). The reported figures should therefore be interpreted as evidence of a clear and consistent direction of effect rather than as precise population estimates, and the Bloom result in particular should be treated as an upper bound until a leakage-controlled re-analysis is available.

6. Conclusion and Outlook

This project set out to close a structural gap: no existing standard combines pedagogical depth with machine-readability. Following the Design Science Research methodology, it developed and evaluated a three-layer best-practice standard for course descriptions that is dual-readable (PDF and JSON-LD) and encodes learning outcomes by Bloom level. Answering the research questions: a compact mandatory-field catalogue with a separated extension layer is sufficient to be both meaningful and machine-processable (RQ1); Bloom's Taxonomy and Constructive Alignment can be operationalised through verb-anchored Bloom levels and an explicit outcome–method–assessment linkage (RQ2); structuring measurably improves extraction and Bloom classification and reduces hallucination (RQ3); and practitioners rate the standard highly on clarity and pedagogical depth while flagging authoring effort as the main barrier (RQ4).

Three lessons learned shaped the work and inform its continuation: a full systematic review is considerably more demanding than an ordinary literature search; understanding the problem thoroughly before designing solutions paid off; and expert interviews need to be scheduled with more lead time than initially planned. The project also confirmed the value of a shared reference manager for collaborative academic work.

Future work follows directly from the limitations, and three steps have priority. First, the AI experiment should be repeated with an evaluation design that separates retrieval from classification: the Bloom answer key should be coded by two independent raters on the original outcome statements before any format is produced, inter-rater agreement should be reported, and Bloom accuracy should be reported separately for the subset of outcomes whose cognitive level is not stated explicitly in any input format. Second, the evaluation should be extended to a larger set of modules and to repeated runs per model–format–task combination, so that means, dispersion and significance can be reported instead of single point values. Third, the revised best-practice artifact (V2) should be validated on these additional modules to strengthen external validity, the expert base should be widened beyond a single institution, and a pilot with EUonAIR partners should test the standard in real recognition workflows. Reducing the token overhead of the JSON-LD representation while preserving its semantic richness is a further worthwhile direction, since it would lower the adoption cost identified in this study.

ISSN 2991-9371

References

[1] Tapia-León, M. et al. (2018): The Use of Ontologies for Syllabus Representation. In: ITHET 2018, pp 1–7.

[2] Vu, B. et al. (2025): InformatIQ: a transformers-powered framework for practical module handbook metadata extraction. In: International Journal of Computer Applications 47 (7), pp 593–610.

[3] Sparmann, S. et al. (2025): Surfacing Educational Traditions in European K–12 Computing Curricula: A Large-Scale NLP-Based Comparison. In: Koli Calling 2025, pp 1–12.

[4] Orellana, G. et al. (2018): A Text Mining Methodology to Discover Syllabi Similarities among Higher Education Institutions. In: INCISCOS 2018, pp 261–268.

[5] Palacio, J. C. C. et al. (2026): An Analysis of ECTS Credit Transferability across Similar Engineering Degree Programs in Spain. In: INTED2026, paper 0958.

[6] Zervas, P. et al. (2014): A Systematic Analysis of Metadata Application Profiles of Learning Object Repositories in Europe. In: LNET, pp 75–91.

[7] Böhm, K. (2021): Towards Semantically Enriched Curricula as pre-built Information Spaces in Higher Education Institutions. In: ECKM 2021, p 71.

[8] Biggs, J. (1996): Enhancing teaching through constructive alignment. In: Higher Education 32 (3), pp 347–364.

[9] Li, Y. et al. (2022): Automatic Classification of Learning Objectives Based on Bloom's Taxonomy. In: Proceedings of EDM 2022.

[10] Kumar, R. et al. (2025): Automated Analysis of Learning Outcomes and Exam Questions Based on Bloom's Taxonomy. arXiv:2511.10903.

[11] Ramesh, R. et al. (2016): Integrating the Learning Objectives and Syllabus into a Domain Ontology for Data Structures Course. In: ITiCSE 2016, pp 266–271.

[12] Liu, Y. et al. (2024): Generative AI in Education: A Review of Applications, Impacts, and Future Research Directions. In: ICIS 2024 Proceedings.

[13] Zaki, N. et al. (2023): Automating the mapping of course learning outcomes to program learning outcomes using natural language processing for accurate educational program evaluation. In: Education and Information Technologies 28 (12), pp 16723–16742.

[14] Krishnan, S. et al. (2025): AI Course Equivalency using Word Mover's Distance and Semantic Analysis with NLP in Cloud Computing. In: ICAISS 2025, pp 448–454.

[15] Webster, J. / Watson, R. T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In: MIS Quarterly 26 (2), pp xiii–xxiii.

[16] Peffers, K. et al. (2007): A Design Science Research Methodology for Information Systems Research. In: Journal of Management Information Systems 24 (3), pp 45–77.

[17] Mayring, P. (2015): Qualitative Inhaltsanalyse. Grundlagen und Techniken. 12th ed., Weinheim: Beltz.

[18] Hevner, A. R. et al. (2004): Design Science in Information Systems Research. In: MIS Quarterly 28 (1), pp 75–105.

[19] Badawy, M. et al. (2018): A Text Mining Approach for Automatic Selection of Academic Course Topics based on Course Specifications. In: ICENCO 2018, pp 162–167.

ISSN 2991-9371

[20] Diab, S. / Sartawi, B. (2017): Classification of Questions and Learning Outcome Statements (LOS) into Bloom's Taxonomy (BT) by Similarity Measurements towards Extracting of Learning Outcome from Learning Material. In: International Journal of Managing Information Technology 9 (2), pp 1–12.

[21] Göschlberger, B. et al. (2020): A Standardised Metadata Schema for Learning Resources based on JSON Schema. In: iiWAS 2020.

[22] Han, Z. / Stevens, K. A. (2026): Quantifying Cross-System Curriculum Alignment: A Reproducible and Interpretable Method. In: SoutheastCon 2026, pp 1–7.

[23] Nilsson, M. (2010): From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization. PhD thesis, KTH Stockholm.

THE USE OF AI TOOLS

The authors used Claude (Anthropic) for language editing and translation of selected parts of the manuscript. All content was reviewed and verified by the authors.